Adaptive Bayes Test For Monotonicity

Jean-Bernard Salomond [*]
CREST and Université Paris Dauphine[†]

March 27, 2013

Abstract

We study the asymptotic behaviour of a Bayesian nonparametric test of qualitative hypotheses. More precisely, we focus on the problem of testing monotonicity of a regression function. Even if some results are known in the frequentist framework, no Bayesian testing procedure has been proposed, at least none has been studied theoretically. This paper propose a procedure that is straightforward to implement, which is a great advantage compared to those proposed in the literature. We describe theoretical properties of this procedure and illustrate its behaviour using a simulation study and real data analysis.

Keywords: Bayesian Nonparametric, Nonparametric regression, Nonparametric hypothesis testing, Asymptotic properties.

1 Introduction

Shape constrained models are of growing interest in the non parametric field. Among them monotonicity constrains are very popular. There is a wide literature on the problem of estimating a function under monotonicity constrains. Groeneboom (1985), Prakasa Rao (1970) and Robertson et al. (1988) among others study the non parametric maximum likelihood estimator of monotone densities, Lo (1984), Brunner and Lo (1989), Khazaei et al. (2012) and Salomond (2013) study the properties of a Bayesian estimator. Barlow et al. (1972) and Mukerjee (1988) proposed a shape constrain estimator of monotonic regression functions. These methods are widely applied in practice. Bornkamp and Ickstadt (2009) consider monotone function when modeling the response to a drug as a function of the dose and Neittaanmäki et al. (2008) use a monotone representation for environmental data.

In this paper, we propose a procedure to test for monotonicity constrains. We consider the Gaussian regression model

$$Y_i = f(i/n) + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right), \sigma^2 > 0,\ i = 1, \dots, n, \tag{1}$$

[*] email : jean.bernard.salomond@ensae.fr

[†] CREST, 3 avenue Pierre Larousse, 92245 Malakoff France.

and, with $\mathcal{F}$ being the set of all monotone function, we test

$$H_0 : f \in \mathcal{F}, \text{ versus } H_1 : f \notin \mathcal{F}. \tag{2}$$

Thus both the null and the alternative are non parametric hypotheses. The problem of testing for monotonicity has already been addressed in the frequentist literature and a variety of approaches have been considered. Baraud et al. (2005) use projections of the regression function on the sets of piecewise constant function on a collection of partition of support of f. Their test rejects monotonicity if there is at least one partition such that the estimated projection is too far from the set of monotone functions. Another approach, considered in Hall and Heckman (2000) and Ghosal et al. (2000) among others, is to test for negativity of the derivative of the regression function. However this requires some assumptions on the regularity of the regression function under the null hypothesis that could be avoided. In a recent paper Akakpo et al. (2012) propose a procedure that detects local departure from monotonicity, and study very precisely its asymptotic properties.

Here, we consider a Bayesian approach to this problem, which to the author's knowledge has not been studied. We only consider the case where $\mathcal{F}$ is the set of monotone non increasing functions, but a similar approach could be used when considering the set of monotone increasing or simply monotone functions. The most common approach to testing in a Bayesian setting is the Bayes Factor. Here however, we see that this method has drawbacks and seems to have poor performances.

1.1 The Bayes Factor Approach

Since monotone non increasing densities are well approximated by piecewise constants, see Groeneboom (1985) or Salomond (2013), it is natural to build a prior on such functions. The standard approach to test for monotonicity of f in a Bayesian setting would be to consider the Bayes Factor

$$B_{0,1} = \frac{\pi\left(f \in \mathcal{F}|Y^n\right)}{\pi\left(f \notin \mathcal{F}|Y^n\right)} \frac{1-\pi\left(\mathcal{F}\right)}{\pi\left(\mathcal{F}\right)}$$

where π has the form, for all $k \geq 2$ and all f written as

$$f = \sum_{i=1}^{k} \mathbb{I}_{[(i-1)/k, i/k)} \omega_i, \; d\pi(f) = \pi(k)\pi(\omega_1, \dots, \omega_k|k).$$

However, this approach seems to lead to poor results in practice. The reason behind this is that when f has flat parts, it becomes difficult to detect monotonicity due to estimation uncertainty. For instance when considering the function $f = 0$ the Bayes Factor does not seem to give a credible answer. As an illustration, Figure 1 gives the histogram constructed from 100 draws of data with $f = 0$ and $n = 100$. It appears that for these runs, the Bayes Factor is rather small and that for a non negligible proportion of samples the log Bayes Factor is negative. Thus the answers given by the Bayes Factor are not satisfying in this case.

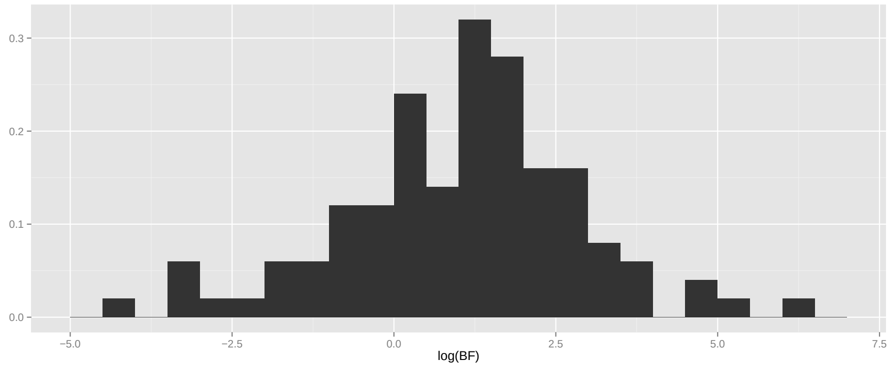

Figure 1: 100 simulation of the log Bayes Factor $B_{0,1}$ for $f = 0$ and $n = 100$

1.2 An alternative approach

To tackle this issue of constructing a test robust to flat parts, we change the formulation of our test into

$$H_0^a : \tilde{d}(f, \mathcal{F}) \leq \tau \;\text{ versus }\; H_1^a : \tilde{d}(f, \mathcal{F}) > \tau \tag{3}$$

where $\tilde{d}(f, \mathcal{F})$ is a distance between f and the set of monotone non increasing function and τ a threshold. To perform such a test we consider the $\gamma_0 - \gamma_1$ loss with fixed $\gamma_0, \gamma_1 > 0$ and thus our procedure can be define as

$$\delta_n^\pi := \begin{cases} 0 \text{ if } \pi\left(\tilde{d}(f, \mathcal{F}) \leq \tau | X_n\right) \geq \frac{\gamma_0}{\gamma_0 + \gamma_1} \\ 1 \text{ otherwise} \end{cases} \tag{4}$$

This ideas is similar to the one proposed in Rousseau (2007) for the approximation of a point null hypothesis by an interval hypothesis testing, see also Verdinelli and Wasserman (1998). The threshold can be calibrated a priori by a prior knowledge on the tolerance to approximate monotonicity. In practice such an a priori calibration is not always feasible. We therefore propose in this paper an automatic calibration of τ. The idea of this construction is to choose τ small enough such that the power of the test is not too deteriorated still remaining robustness to flat parts under the null. The resulting procedure has good asymptotic properties, see Theorem 1, but behave also well in finite sample situations, as shown in section 3. Furthermore form a practical point on view, this procedure will be easy to implement as it will only require sampling under the posterior distribution. This is a great advantage compare to the frequentist tests proposed in the literature as they require in general heavy computations.

We calibrate τ such that our test is consistent, that is for all $\rho > 0$ and $d(\cdot, \cdot)$

a metric

$$\begin{aligned} \sup_{f \in \mathcal{F}} \mathrm{E}_0^n(\delta_n^\pi) = o(1) \\ \sup_{f, d(f,\mathcal{F}) > \rho} \mathrm{E}_0^n(1 - \delta_n^\pi) = o(1). \end{aligned} \tag{5}$$

where $d(f, \mathcal{F}) := \inf_{g \in \mathcal{F}} d(f, g)$. In absence of prior information on the threshold, it is natural to have τ depending on n, since the more data, the more precise we can afford to be. Hence, to understand better the effectiveness of the threshold induced by our approach, we study the minimum separation rate of our test which is the minimum value $\rho = \rho_n$ such that (5) is still valid. Small ρ_n implies that the test is able to detect very small departure from the null. We want our calibrated threshold to induce the smallest separation rate.

We thus propose a procedure which although being a Bayesian answer to the problem (3), is also asymptotically an answer to the problem (2). Moreover, our procedure is automatic and easy to implement. The construction of the test is presented in section 2 and its asymptotic properties are discussed in Section 2.2. In Section 2.3 we propose a way to calibrate the hyperparameters of the prior rending the procedure fully automatic. We then run our test on simulated data in section 3 and on real environmental data in section 4. A general discussion is provided in section 5.

2 Construction of the test

In this section we present our testing procedure based on the approach described in section 1.2. We propose a choice for $\tilde{d}(f, \mathcal{F})$ which measures the distance between the regression function f and the set $\mathcal{F}$. We also give an autocalibrated threshold τ such that by answering the problem (3) we give a good answer to the problem (2). We then propose a specific family of prior together with a choice for the hyperparameters based on heuristics.

2.1 The testing procedure

As presented in section 1, monotone non increasing functions are well approximated by stepwise constant functions. Let $\mathcal{G}_k$ be the set of piecewise constant function with k pieces on the partition $\{[0, 1/k), \dots [(k-1)/k, 1]\}$. We denote $f_{\omega,k} \in \mathcal{G}_k$ the function

$$f_{\omega,k}(\cdot) = \sum_{i=1}^{k} \omega_i \mathbb{I}_{[(i-1)/n, i/n)}(\cdot). \tag{6}$$

In model (1), we consider the residual variance σ^2 to be unknown. We then build a prior on (f, σ) taking a prior on k and building a prior on each submodels $\mathcal{G}_k$. We define

$$\pi(\omega, \sigma, k) := \pi(k)\pi(\sigma|k)\pi(\omega|\sigma, k)$$

First note that with this choice of prior we have generally speaking $\pi(\mathcal{F}) > 0$. Furthermore, if the true regression function f_0 is in $\mathcal{F}$ then the piecewise constant function of the form (6) which minimize the Kullback Leibler divergence with f_0 will also be in $\mathcal{F}$. We consider the following discrepancy measure $\tilde{d}(\cdot,\cdot)$

$$\tilde{d}(f_{\omega,k}, \mathcal{F}) = H(\omega, k) = \max_{k \geq j \geq i \geq 1} (\omega_j - \omega_i) \tag{7}$$

From (7) it appears that $f_{\omega,k}$ is in $\mathcal{F}$ if and only if $\tilde{d}(f_{\omega,k}, \mathcal{F}) = 0$. Here the discrepancy $\tilde{d}$ corresponds to the sup norm between $f_{\omega,k}$ and the set of monotone non increasing functions. The idea of the calibration is the following. In the model $\mathcal{G}_k$, the a posteriori uncertainty for estimating $\omega = (\omega_1, \dots, \omega_k)$ is of order $\sqrt{k/n}$. Hence any monotone non increasing function $f_{\omega,k}$ such that for all $j > i$, $\omega_i \geq \omega_j - O(\sqrt{k/n})$ might be detected as possibly monotone non increasing. We thus decide to construct a threshold τ_n^k for each model $\mathcal{G}_k$. We then compare $H(\omega, k)$ with some positive threshold depending on n and k. We then calibrate τ_n^k such that our procedure is consistent. Similarly to the frequentist procedures, we consider Hölderian alternatives

$$f \in \mathcal{H}(\alpha, L) = \left\{ f, [0,1] \to \mathbb{R}, \forall x, y \in [0,1]^2 |f(y) - f(x)| \leq L|y - x|^\alpha \right\}$$

for some constant $L > 0$ and a regularity parameter $\alpha \in (0, 1]$. We study the separation rate of our procedure and compare it with the minimax separation rate $n^{-\alpha/(2\alpha+1)}$.

2.2 Theoretical results

The following Theorem gives a calibration for τ_n^k. It also gives an upper bound for the minimal separation rate with respect to the distance $d_n(\cdot,\cdot)$ defined as

$$d_n^2(f, g) = n^{-1} \sum_{i=0}^{n-1} \left(f\left(\frac{i}{n}\right) - g\left(\frac{i}{n}\right) \right)^2.$$

We define a prior π on f, σ similarly to before by considering

$$f_{\omega,k}(\cdot) = \sum_{i=1}^{k} \omega_i \mathbb{I}[(i-1)/k, i/k)(\cdot)$$

and

$$d\pi(\omega, \sigma, k) = \pi_k(k) \pi_\sigma(\sigma) \prod_{i=1}^{k} g(\omega_i)$$

where g and π_σ are density function. We consider the following conditions on the prior

C1 The densities g and π_σ are continuous, $g(x) > 0$ for all $x \in \mathbb{R}$ and $\pi_\sigma(\sigma) > 0$ for all $\sigma \in (0, \infty)$.

C2 π_k is such that there exists positive constants C_d and C_u such that

$$e^{-C_d k L(k)} \leq \pi_k(k) \leq e^{-C_u k L(k)} \tag{8}$$

where $L(k)$ is either $\log(k)$ or 1.

The condition **C1** is mild as it is satisfied for a large variety of distributions. **C2** is an usual condition when considering mixture models with random number of components (see e.g. Rousseau (2010)) and is satisfied by Poisson or Geometric distribution for instance. Under this conditions, Theorem 1 gives us some insight on how to choose τ_n^k.

Theorem 1 *Under the assumptions **C1** and **C2**, if $M_0 > 0$, setting $\tau = \tau_n^k = M_0\sqrt{k\log(n)/n}$ and δ_n^π the testing procedure*

$$\delta_n^\pi = \mathbb{I}\left\{\pi\left(H(\omega,k) > \tau_n^k | Y^n\right) > \gamma_0/(\gamma_0+\gamma_1)\right\}$$

then there exists some $M > 0$ such that for all $\alpha \in (0,1]$

$$\begin{aligned}\sup_{f\in\mathcal{F}} \mathrm{E}_f^n(\delta_n^\pi) &= o(1)\\ \sup_{f, d_n(f,\mathcal{F})>\rho, f\in\mathcal{H}(\alpha,L)} \mathrm{E}_f^n(1-\delta_n^\pi) &= o(1)\end{aligned} \tag{9}$$

for all $\rho > \rho_n = M(n/\log(n))^{-\alpha/(2\alpha+1)}$.

Note that neither the prior nor the hyperparameters depend on the regularity α of the regression function under the alternative. Moreover for all $\alpha \in (0,1]$, the separation rate $\rho_n(\alpha)$ is the minimax separation rate up to a $\log(n)$ term. Thus our test is almost minimax adaptive. The $\log(n)$ term seems to follow from our definition of the consistency where we do not fix a level for the Type I or Type II error contrariwise to the frequentist procedures. The conditions on the prior are quite loose, and are satisfied in a wide variety of cases. The constant M_0 does not influence the asymptotic behaviour of our test but has a great influence in practice for finite n. A way of choosing M_0 is given in section 2.3.

The proof of Theorem 1 is given in Appendix A, we now sketch the main ideas. We approximate the true regression function f_0 in each submodel $\mathcal{G}_k$ of piecewise constant functions associated with k pieces on $\{[0,1/k),\dots[1-1/k,1)\}$ by $f_{\omega^0,k}$ by $f_{\omega^0,k}$ that minimize the Kullback-Leibler divergence with f_0. We get a close form expression for $\omega^0 = (\omega_1^0,\dots,\omega_k^0)$ defined by

$$\omega_i^0 = n_i^{-1}\sum_{j,j/n\in[(i-1)/k,i/k)} f_0(j/n),\ n_i = \mathrm{Card}\left\{j, j/n \in [(i-1)/k, i/k)\right\} \tag{10}$$

thus $f_{\omega^0,k}$ belongs to $\mathcal{F}$ for all k when $f_0 \in \mathcal{F}$. To prove the first part of (9), we bound $H(\omega,k) \leq 2\max|\omega_i - \omega_i^0|$ if $f_0 \in \mathcal{F}$ so that the threshold τ_n^k needs to be as large as the posterior concentration rate of ω to ω^0 in the misspecified model $\mathcal{G}_k$. Then to prove the second part of (9) when $\rho = \rho_n(\alpha)$, we bound form below $H(\omega,k)$ by $H(\omega^0,k) - 2\max|\omega_i - \omega_i^0|$ which implies a constraint on the separation rate of the test to ensure that uniformly over $d_n(f_0,\mathcal{F}) \geq \rho_n(\alpha)$ and $f \in \mathcal{H}(\alpha,L)$ we have $H(\omega,k) > \tau_n^k$.

2.3 A choice for the prior in the non informative case

Conditions on the prior in Theorem 1 are satisfied for a wide variety of distributions. However, when no further informations are available, some specific choices can ease the computations and lead to good results in practice. We present in this section such a specific choice for π and a way to calibrate the hyperparameters. We also fix $\gamma_0 = \gamma_1 = 1/2$ in the definition of δ_n^π.

A practical default choice is the usual conjugate prior, given k, i.e. a Gaussian prior on ω with variance proportional to σ^2 and an Inverse Gamma prior on σ^2. This will considerably accelerate the computations as sampling under the posterior is then straightforward. Condition (8) on π_k is satisfied by the two classical distributions on the number of parameters in a mixture model, namely the Poisson distribution and the Geometric distribution. It seems that choosing a Geometric distribution is more appropriate as it is less spiked. We thus choose

$$\pi := \begin{cases} k \sim \text{Geom}(\lambda) \\ \sigma^2|k \sim IG(a,b) \\ \omega_i|k,\sigma \overset{iid}{\sim} \mathcal{N}(m, \sigma^2/\mu) \end{cases} \tag{11}$$

Standard algebra leads to a close form for the posterior distribution up to a normalizing constant. Denoting $n_j = \text{Card}\,\{i, i/n \in [(j-1)/k, j/k)\}$ and

$$\tilde{b}_k = b + \frac{1}{2}\sum_{j=1}^{k}\left\{\sum_{i,i/n\in I_j}\big(Y_i - \overline{Y_j}\big)^2 + \frac{n_j\mu}{n_j+\mu}(\overline{Y_j}-m)^2\right\},$$

where $\overline{Y_j}$ is the empirical mean of the Y_l on the set $\{l, l/n \in [(j-1)/n, j/n)\}$, we have

$$\pi_k(k|Y^n) \propto \pi(k)\tilde{b}_k^{-(\alpha+n/2)}\mu^{k/2}\prod_{j=1}^{k}(n_j+\mu)^{-1/2}$$

We can thus compute the posterior distribution of k up to a constant. To sample from π_k we will use a random walk Hasting-Metropolis algorithm. We then compute the posterior distribution of ω and σ given k

$$\begin{aligned} &\sigma^2|k, Y^n \sim IG(a+n/2, \tilde{b}_k) \\ &\omega_j|k,\sigma^2,Y^n \overset{ind.}{\sim} \mathcal{N}\left(\frac{m\mu + n_j\bar{Y}_j}{n_j+\mu}, \frac{\sigma^2}{n_j+\mu}\right). \end{aligned}$$

Note that given k, sampling from the posterior is straightforward. We now give a way to calibrate the hyperparameters a, b, μ, m, M_0. We first calibrate M_0 the constant in τ_n^k. The most difficult function in $\mathcal{F}$ to be detected as belonging to $\mathcal{F}$ are the constant functions. We calibrate M_0 by choosing the smallest value such that $\mathrm{E}_f^n(\delta_n^\pi) \leq \alpha$ for $f = 0$ and for reasonable sample size n. Using the fact that the ω_i are a posteriori Gaussian and denoting $z_n = \frac{1}{1+k\sigma^2/n}$ we have, assuming that $\forall j, n_i = n/k$,

$$\pi\Big(H(\omega,k) \geq \tau_n^k|Y^n,k,\sigma\Big) = \pi\Big(\max_{1\leq i<j\leq k}\big(\frac{\bar{Y}_j - \bar{Y}_i}{z_n} + \sqrt{\frac{k}{nz_n}}(U_j - U_i)\big) \geq \tau_n^k|Y^n,k,\sigma\Big)$$

where $U_i \overset{iid}{\sim} \mathcal{N}(0,1)$. This implies that

$$\pi\Big(H(\omega,k) \geq \tau_n^k | Y^n, k, \sigma\Big) \leq \sum_{1\leq i<j\leq k} 1 - \Phi\left(\sqrt{nz_n/2k}\Big(\tau_n^k - \frac{\bar{Y}_j - \bar{Y}_i}{z_n}\Big)\right)$$
$$= \sum_{1\leq i<j\leq k} \Phi\left(\sqrt{nz_n/2k}\Big(\frac{\bar{Y}_j - \bar{Y}_i}{z_n} - \tau_n^k\Big)\right)$$

Given that $f = 0$ we have that $\bar{Y}_i \sim \mathcal{N}(0, n\sigma^2/k)$. In this case we can easily prove that when $f = 0$, $\pi(k = 2|Y^n) = 1 + o_{P_0^n}(1)$ (using the same approach as in Lemma 3 given in Appendix A). Thus restricting our attention to $k = 2$ and σ close to σ_0, we have

$$E_f^n(\delta_n^\pi) \leq P_f^n\left(\Phi\left(\sqrt{nz_n/2k}\Big(\frac{\bar{Y}_j - \bar{Y}_i}{z_n} - \tau_n^k\Big)\right) \geq 1/2\right)$$
$$= P_f^n\left(\Big(\frac{\bar{Y}_j - \bar{Y}_i}{z_n} \geq M_0\sqrt{\frac{\log(n)k}{n}}\Big)\right)$$

and thus choose

$$M_0 = \frac{\Phi^{-1}(1-\alpha)\sqrt{2}\hat{\sigma}}{\sqrt{\log(n)}}$$

where $\hat{\sigma}$ is the posterior mean of $\sigma|k, Y^n$. And thus have $\mathrm{E}_{f=0}^n(\delta_n^\pi) \leq \alpha$.

We now propose a calibration for a, b, m, μ and λ. We first choose m to be the empirical mean of the Y_i. We then chose a and b such that the prior on σ has a first order moment and $\mathrm{E}_\pi(\sigma^2)$ is of the same order as σ_y^2. We choose $a = \sigma_y^2 + 1$ and $b = \sigma_y^4$. We want the prior on ω to be flat enough to recover large variation from the mean m. This is done by choosing the hyperparameter μ small enough. We also want the prior on k to be flat to allow large values of k even for small samples sizes. We calibrate μ and λ on simulated data when $f = 0$ and $\sigma = 1$, and choose the minimum values such that $E_0^n(\delta_n^\pi) \leq 0.05$.

3 Simulated Examples

In this section we run our testing procedure on simulated data to study the behaviour of our test for finite sample size. We choose the prior distribution and calibrate the hyperparameters as exposed in section 2.3. We consider nine functions adapted from Baraud et al. (2003) and plot in Figure 2.

$$\begin{aligned}
f_1(x) =& -15(x-0.5)^3\mathbb{I}_{x\leq 1/2} - 0.3(x-0.5) + e^{-250(x-0.25)^2}\\
f_2(x) =& 0.15x\\
f_3(x) =& 0.2e^{-50(x-0.5)^2}\\
f_4(x) =& -0.5\cos(6\pi x)\\
f_5(x) =& -0.2x + f_3(x)\\
f_6(x) =& -0.2x + f_4(x)\\
f_7(x) =& -(1+x) + 0.25e^{-50(x-0.5)^2}\\
f_8(x) =& -0.5x^2\\
f_9(x) =& 0
\end{aligned} \tag{12}$$

The functions f_1 to f_6 are clearly not in $\mathcal{F}$. The function f_7 has a small bump around $x = 0.5$ which can be seen as a local departure from monotonicity. This function is thus expected to be difficult to detect for small datasets given our parametrization. The function f_9 is a completely flat function which is the most difficult situation under H_0.

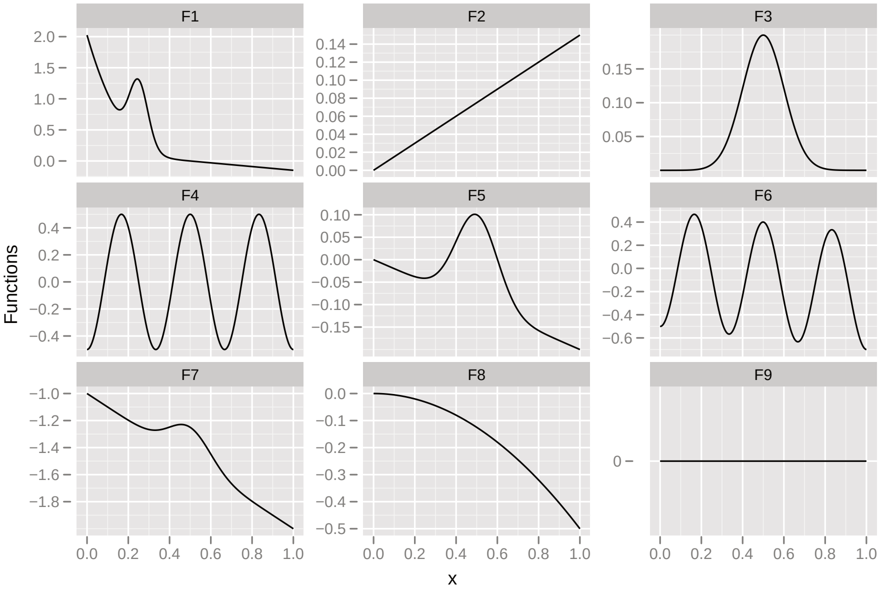

Figure 2: Regression functions used in the simulated example.

For several values of n, we generate $N = 500$ replication of the data $Y^n = \{y_i, i = 1 \dots n\}$ from model (1). For each replication we draw $K = 5.10^3$ iterations from the posterior distribution using a Hasting-Metropolis sampler

Table 1: Percentage of rejection for the simulated examples

	f_0	σ^2	Barraud et	Akakpo et	Bayes Test, n :				
			al. $n = 100$	al. $n = 100$	100	250	500	1000	2500
H_0	f_1	0.01	99	99	100	100	100	100	100
	f_2	0.01	99	100	98.4	100	100	100	100
	f_3	0.01	99	98	98.8	100	100	100	100
	f_4	0.01	100	99	100	100	100	100	100
	f_5	0.004	99	99	99.5	100	100	100	100
	f_6	0.006	98	99	100	100	100	100	100
	f_7	0.01	76	68	18.6	41.8	66.6	86.4	100
H_1	f_8	0.01	-	-	2.4	1.8	1.6	1.4	3.6
	f_9	0.01	-	-	5.6	5.1	5.2	5.1	4.2

with a compound Geometric proposal. More precisely, if k_{i-1} the state of our Markov chain at the step i, we propose

$$k_i^p = k_{i-1} + p_i$$

where p_i is such that

$$|p_i| \sim \text{Geom}(0.3) + 1$$
$$P(p_i < 0) = P(p_i > 0) = \frac{1}{2}$$

Given k we draw directly σ^2 and ω from the marginal posteriors. We then approximate $\pi\left(H(\omega, k) > \tau_n^k | Y^n\right)$ by the standard Monte Carlo estimate

$$\hat{\pi}\left(H(\omega, k) > \tau_n^k | Y^n\right) = \frac{1}{K}\sum_{i=1}^{K} \mathbb{I}\left\{H(\omega^i, k^i) > \tau_n^{k^i}\right\}$$

and reject the null if $\hat{\pi}\left(H(\omega, k) > \tau_n^k | Y^n\right) > 1/2$. The results are given in table 1.

For all the considered functions, the computational time is reasonable even for large values of n. For instance, for f_1, we require less than 25 seconds to perform the test for $n = 2500$ using a simple Python script available on the author's webpage. For the models with regression function f_1 to f_7, we choose the same residuals variance as in Baraud et al. (2003), for the last two functions, we choose a variance of 0.01 which is of the same order. We observe that for the regression functions f_1 to f_6, the test perform well and reject monotonicity for almost all tested samples even when n is small. The results obtained for $n = 100$ are comparable with those obtained in Akakpo et al. (2012) and Baraud et al. (2003). We observe a consequent loss of power for f_7 for small sample size. For this last function, the deviation from monotonicity is small and local, making it hard to detect using piecewise constant functions. In fact, computing $H(\omega^0, k)$ for this function for different values of k shows that we need $k > 13$ get $H(\omega^0, k) > 0$ and thus detect departure from monotonicity. However with our choice of hyperparameters, the posterior distribution puts most of its mass on small values of k when n is small.

It thus appears that our procedure requires a larger amount of data to detect local departure from monotonicity than the frequentist ones. Nonetheless, our test is easy to run on large datasets and has good performances when $n \geq 1000$.

4 Application to Global Warming data

We consider the Global Warming dataset provided by Jones et al. (2011) plotted in Figure 4. It contains the annual temperatures anomalies from 1850 to 2010, expressed in degrees Celcius. Temperature anomaly is the departure from a long-term average, here the 1961-1990 mean. The data are gathered from both land and sea meteorological stations and corrected for non climatic error. In the literature, this dataset has been used to illustrate some isotonic regression techniques in Wu et al. (2001) and Zhao and Woodroofe (2012) where they use frequentist estimation procedures under monotonicity constraint. Alvarez and Dey (2009) show, using a Bayesian monotonic change point method, that there is a positive trend, and that this trend tend to increase of about $.3°C$ in the global annual temperature between 1958 and 2000. Álvarez and Yohai (2012) show that the phenomenon of global warming is due to a steady increase trend phenomenon using a isotonic estimation methods. In our model, that would mean that the regression function f should be positive increasing and convexe. In all these papers the data is supposed to be a sequence of independent and identically distributes random variables. This assumption is questionable (see Fomby and Vogelsang (2002)), but considering annual temperature anomalies should reduce the serial correlation. Similarly to these authors, we make the same assumption of independence. Our aim is to test if the hypothesis of increasing temperature anomaly is realistic, given the amount of information, using the method described in section 1. In particular, we choose the prior and the hyperparameters based on the rule described in section 2.

We perform our test on this dataset (more precisely on minus the temperature anomalies to test for monotone increasing trend), choosing the hyperparameters as in section 2.3. We run the MCMC sampler described above for $K = 10^5$ in order to compute Monte Carlo estimate of δ_n^π. We obtained

$$\hat{\pi}(H(\omega, k) > \tau_n^k | Y^n) = 0.98$$

and thus the hypothesis of monotony is ruled out by our procedure. We conclude that applying a shape constrained regression techniques on the trend of this dataset can deteriorate the estimation results.

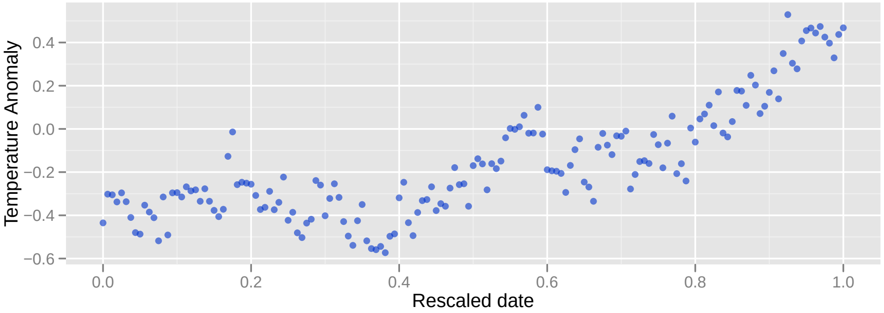

Figure 3: Plot of the Global Warming data

5 Discussion

In this paper we propose a Bayesian approach to the problem of testing qualitative hypotheses in a non parametric framework. More precisely we address the problem of testing monotonicity of a regression function. This problem arise naturally as shape constrained models, and monotonicity in particular, are fairly used in practice. Our approach is particularly interesting as it focuses on a problem where the Bayes Factor seems to give poor results and thus an alternative approach should be considered.

The testing procedure proposed in this paper is a modified version of the Bayes Factor that only reject H_0 when the data gives strong evidence that the function is not monotone. When possible, one can choose a threshold based on prior information on the tolerance level to non monotony. However, this could be difficult in practice, we thus present a way to calibrate our test such that it behave well asymptotically. Interestingly this calibration leads to the optimal separation rate (up to a $\log(n)$ term) and thus the tolerance induced by our approach, and the fact that we test (3) (H_0^a versus H_1^a) instead of (2) (H_0 versus H_1) is of the same order as the classical tests available in the literature. It has the advantage of being very simple to implement even in presence of large datasets. Although we have focused on monotonicity constrains, other types of shape constrains such as convexity or unimodality can be dealt with using this approach. For instance we can test for convexity using piecewise linear functions as submodels $\mathcal{G}_k$ and test monotonicity of the slope.

References

Akakpo, N., Balabdaoui, F., and Durot, C. (2012). Testing monotonicity via local least concave majorants.

Alvarez, E. E. and Dey, D. K. (2009). Bayesian isotonic changepoint analysis. *Ann. Inst. Statist. Math.*, 61(2):355–370.

Álvarez, E. E. and Yohai, V. J. (2012). M-estimators for isotonic regression. *J. Statist. Plann. Inference*, 142(8):2351–2368.

Baraud, Y., Huet, S., and Laurent, B. (2003). Adaptive tests of qualitative hypotheses. *ESAIM Probab. Stat.*, 7:147–159.

Baraud, Y., Huet, S., and Laurent, B. (2005). Testing convex hypotheses on the mean of a Gaussian vector. Application to testing qualitative hypotheses on a regression function. *Ann. Statist.*, 33(1):214–257.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical inference under order restrictions. The theory and application of isotonic regression*. John Wiley & Sons, London-New York-Sydney. Wiley Series in Probability and Mathematical Statistics.

Bornkamp, B. and Ickstadt, K. (2009). Bayesian nonparametric estimation of continuous monotone functions with applications to dose-response analysis. *Biometrics*, 65(1):198–205.

Brunner, L. J. and Lo, A. Y. (1989). Bayes methods for a symmetric unimodal density and its mode. *Ann. Statist.*, 17(4):1550–1566.

Choi, T. and Schervish, M. J. (2007). On posterior consistency in nonparametric regression problems. *J. Multivariate Anal.*, 98(10):1969–1987.

Fomby, T. B. and Vogelsang, T. J. (2002). The Application of Size-Robust Trend Statistics to Global-Warming Temperature Series. *Journal of Climate*, 15:117–123.

Ghosal, S., Sen, A., and van der Vaart, A. W. (2000). Testing monotonicity of regression. *Ann. Statist.*, 28(4):1054–1082.

Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.*, 35(1):192–223.

Groeneboom, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, Wadsworth Statist./Probab. Ser., pages 539–555, Belmont, CA. Wadsworth.

Hall, P. and Heckman, N. E. (2000). Testing for monotonicity of a regression mean by calibrating for linear functions. *Ann. Statist.*, 28(1):20–39.

Jones, P., Parker, D., Osborn, T., , and Briffa, K. (2011). Global and hemispheric temperature anomalies, land and marine instrumental records.

Khazaei, S., Rousseau, J., and Balabdaoui, F. (2012). Nonparametric bayesian estimation of densities under monotonicity constraint.

Kleijn, B. J. K. and van der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.*, 34(2):837–877.

Le Cam, L. (1986). *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. Springer-Verlag, New York.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Stat.*, 12:351–357.

Mukerjee, H. (1988). Monotone nonparameteric regression. *Ann. Statist.*, 16(2):741–750.

Neittaanmäki, P., Rossi, T., Majava, K., and Pironneau, O. (2008). Monotonic regression for assessement of trends in environmental quality data.

Prakasa Rao, B. L. S. (1970). Estimation for distributions with monotone failure rate. *Ann. Math. Statist.*, 41:507–519.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester.

Rousseau, J. (2007). Approximating interval hypothesis: p-values and Bayes factors. In *Bayesian statistics 8*, Oxford Sci. Publ., pages 417–452. Oxford Univ. Press, Oxford.

Rousseau, J. (2010). Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *Ann. Stat.*, 38(1):146–180.

Salomond, J.-B. (2013). Concentration rate and consistency of the posterior under monotonicity constraints. *ArXiv e-prints*.

Verdinelli, I. and Wasserman, L. (1998). Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *Ann. Statist.*, 26(4):1215–1241.

Wu, W. B., Woodroofe, M., and Mentz, G. (2001). Isotonic regression: another look at the changepoint problem. *Biometrika*, 88(3):793–804.

Zhao, O. and Woodroofe, M. (2012). Estimating a monotone trend. *Statist. Sinica*, 22(1):359–378.

A Proof of Theorem 1

In order to prove Theorem 1 we need some concentration results of the posterior around the true regression function. The following Lemma provides a posterior concentration rate when f_0 is either in $\mathcal{F}$ or in $\mathcal{H}(\alpha, L)$. The proof is given in appendix B and is derived from Ghosal and van der Vaart (2007). Some adaptive results are known for the Gaussian regression under some regularity assumptions, the monotone case has not been studied and thus this Lemma has an interest in its own.

Let $d_n(\cdot, \cdot)$ be define as

$$d_n(f,g)^2 = n^{-1} \sum_{i=1}^{n} \left(f(i/n) - g(i/n) \right)^2$$

and denote P_0^n the distribution of the Y_i when $f = f_0$ in (1).

Lemma 1 *Let f_0 be either in $\mathcal{F}$ or in $\mathcal{H}(\alpha, L)$, and let π be defined as in Theorem 1. Thus*

$$\mathrm{E}_{P_0^n} \left(\pi (d_n(f_{\omega,k} - f_0)^2 + (\sigma - \sigma_0)^2 \geq M\epsilon_n^2 | Y^n) \right) \to 0$$

where $\epsilon_n = (n/\log(n))^{1/3}$ if $f_0 \in \mathcal{F}$ and $\epsilon_n = (n/\log(n))^{-\alpha/(2\alpha+1)}$ if $f_0 \in \mathcal{H}(\alpha, L)$.

The proof of this lemma is postponed to Appendix B. Given this result, we get the following Lemma that enable us to derive consistency and an upper bound on the separation rate.

Lemma 2 *Let M be a positive constant and $\rho_n(\alpha) = M(n/\log(n))^{-\alpha/(2\alpha+1)}$. Let π be as in Theorem 1 and ω_0 be the minimizer of the Kulback-Leibler divergence $KL(f_{\omega,k}, f_0)$. Then*

$$\pi\left(\max_i |\omega_i - \omega_i^0| \geq C\xi_n^k | Y^n\right) \leq 1/2 + o_{P_0^n}(1). \tag{13}$$

where ξ_n^k is either equal to τ_n^k when $f_0 \in \mathcal{F}$ or $\rho_n(\alpha)$ when $f_0 \notin \mathcal{F}$ and $f_0 \in \mathcal{H}(\alpha, L)$.

The proof of this lemma is postponed to Appendix B. Given the preceding results, we derive (9).

Consistency under H_0 Let $f_0 \in \mathcal{F}$ then

$$H(\omega, k) \leq 2 \max_i |\omega_i - \omega_i^0|$$

and thus

$$\pi(H(\omega, k) < \tau_n^k | Y_n) < 1/2 + o_{P_0^n}(1)$$

which gives the consistency under H_0 given Lemma 2.

Consistency under H_1 and upper bound for the separation rate Let $f_0 \notin \mathcal{F}$ and $f_0 \in H_\alpha(L)$ we have

$$H(\omega, k) \geq H(\omega_0, k) - 2 \max_i |\omega_i - \omega_i^0| \tag{14}$$

Assume that $\rho_n(\alpha) < d_n(f_0, \mathcal{F})$, we derive a lower bound for $H(\omega^0, k)$. Let g^* be the monotone non increasing piecewise constant function on the partition $\{[0, 1/k), \dots, [(k-1)/k, 1)\}$, with for $1 \leq i \leq k$, $g_i^* = \min_{j \leq i} \omega_j^0$. Given that $d_n(f_{\omega^0,k}, \mathcal{F}) = \inf_{g \in \mathcal{F}} d_n(f_{\omega^0,k}, g)$ we get

$$d_n(f_{\omega^0,k}, \mathcal{F}) \leq d_n(f_{\omega^0,k}, g^*) \leq H(\omega^0, k)$$

And therefore, given that $d_n(f_0, \mathcal{F}) \leq d_n(f_{\omega^0,k}, \mathcal{F}) + d_n(f_{\omega^0,k}, f_0)$

$$\pi\left(H(\omega, k) < C\tau_n^k | Y_n\right) \leq \pi\left(\max_i |\omega_i - \omega_i^0| \geq \frac{\rho_n(\alpha) - d_n(f_{\omega^0,k}, f_0) - C\tau_n^k}{4} | Y^n\right)$$

The following Lemma states that the posterior probability that k is greater that $K_0 n^{1/(2\alpha+1)}$ is less than a $o_{P_0^n}(1)$.

Lemma 3 *Let $\mathcal{K}_n = \{k \leq K_0(n/\log(n))^{1/(2\alpha+1)}\}$. If π is define as in Theorem 1 and $f_0 \in \mathcal{H}(\alpha, L)$, then*

$$\pi\left(\mathcal{K}_n^c | Y^n\right) \leq o_{P_0^n}(1) \tag{15}$$

The proof is postponed to Appendix B

For $k \in \mathcal{K}_n$ and M large enough we have $\rho_n(\alpha)/4 > \tau_n^k$. Denoting $B_n = \{d_n(f_{\omega,k}, f_0)^2 + |\sigma_0 - \sigma|^2 \le \epsilon_n^2\}$, Lemma 1 gives

$$\pi(B_n^c|Y_n) = o_{P_0^n}(1).$$

On the set $B_n \cap \mathcal{K}_n$ we have for R large enough $\rho_n(\alpha)/4 \ge d_n(f_{\omega^0,k}, f_0)$

$$\pi\left(H(\omega,k) < C\tau_n^k|Y_n\right) \le \pi\left(\{\max_i |\omega_i - \omega_i^0| \ge \rho_n(\alpha)/8\} \cap \{\mathcal{K}_n \cap B_n\}|Y^n\right) + o_{P_0^n}(1).$$

Given (13), we get that for all f_0 such that $d_n(f_0, \mathcal{F}) > \rho_n(\alpha)$

$$\pi(H(\omega,k) < \tau_n^k|Y_n) < \frac{1}{2} + o_{P_0^n}(1)$$

which ends the proof.

B Proof of Lemmas 1, 2 and 3

B.1 Proof of Lemma 1

In this section we prove that the posterior concentrate around f_0, σ_0 at the rate $(n/\log(n))^{-1/3}$ if $f_0 \in \mathcal{F}$ and $(n/\log(n))^{-\alpha/(2\alpha+1)}$ if $f_0 \in \mathcal{H}(\alpha, L)$. To do so we follow the approach of Ghosal and van der Vaart (2007). Let $KL(f,g) = \int f \log(f/g)$ be the Kullback-Leibler divergence between the two probability densities f and g. We define $V(f,g) = \int(\log(f/g) - KL(f,g))^2 f$. We denote $P_i(\omega,\sigma,k)$ the probability measure of $Y_i = f_{\omega,k} + \epsilon_i$ and $p_i(\omega,\sigma,k)$ its density with respect to the Lebesgue measure when $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$ and $P_{i,0}$ the true distribution of Y_i and $p_{i,0}$ its density. We only consider the case where $f \in \mathcal{F}$, a similar proof holds when $f \in \mathcal{H}(\alpha, L)$. We define

$$B_n(\epsilon) = \left\{\sum_{i=1}^n KL(p_i(\omega,\sigma,k), p_{i,0}) \le n\epsilon^2, \sum_{i=1}^n V(p_i(\omega,\sigma,k), pi, 0) \le n\epsilon^2\right\}$$

Here $p(\omega,\sigma,k)$ and p_0 are Gaussian distributions, we can easily compute

$$KL(p_i(\omega,\sigma,k), p_{i,0}) = \frac{1}{2}\log\left(\frac{\sigma^2}{\sigma_0^2}\right) - \frac{1}{2}\left(1 - \frac{\sigma_0^2}{\sigma^2}\right) + \frac{1}{2}\frac{(f_{\omega,k}(x_i) - f_0(x_i))^2}{\sigma^2}$$
$$V(p_i(\omega,\sigma,k), pi, 0) = \frac{1}{2}\left(1 - \frac{\sigma_0^2}{\sigma^2}\right)^2 + \left[\frac{\sigma_0^2}{\sigma^2}(f_{\omega,k}(x_i) - f_0(x_i))\right]^2$$

We have $B_n(\epsilon_n) \supset \{d_n^2(f_{\omega,k}, f_0) \le C\epsilon_n, |\sigma^2 - \sigma_0^2|^2 \le C\epsilon_n^2\}$.

For $f_0 \in \mathcal{F}$, denoting $\omega_j^0 = n_j^{-1}\sum_{x_i \in I_j} f_0(x_i)$ and $\underline{x_j} = \inf(I_j)$, $\overline{x_j} = \sup(I_j)$ we have

$$d_n^2(f_{\omega,k}, f_0) = d_n^2(f_0, f_{\omega^0,k}) + d_n^2(f_{\omega,k}, f_{\omega^0,k})$$

and

$$\begin{aligned}
d_n^2(f_0, f_{\omega^0,k}) &= \frac{1}{n}\sum_{j=1}^{k}\sum_{x_i\in I_j}(f_0(x_i) - f_{\omega^0,k})^2 \\
&\leq \frac{1}{n}\sum_{j=1}^{k} n_j(f(\underline{x_j}) - f(\overline{x_j}))^2 \\
&\leq \frac{c}{k}\left(\sum_{j=1}^{k}(f(\underline{x_j}) - f(\overline{x_j}))\right)^2 \leq \frac{c||f_0||_\infty^2}{k}.
\end{aligned}$$

Denoting $k_n = C\lceil(n/\log(n))^{1/3}\rceil$ we deduce that $B_n(\epsilon_n) \supset \{k = k_n, ||\omega - \omega^0||_n^2 \leq \epsilon_n^2, |\sigma^2 - \sigma_0^2| \leq \epsilon_n^2\}$ where $||\cdot||$ is the standard Euclidean norm in $\mathbb{R}^{k_n}$. We deduce that

$$\pi(B_n(\epsilon_n)) \geq \mathrm{C}\left(\inf_{x\in[0,1]}(\pi_\omega(f_0(x)))\epsilon_n\right)^{k_n}\pi_\sigma(\sigma_0^2)\epsilon_n^2\pi(k = k_n) \geq e^{-C_0 n\epsilon_n^2} \quad (16)$$

To end the proof of Lemma 1, the standard approach of Ghosal and van der Vaart (2007) require the existence of an exponentially consistent sequence of tests. Their Theorem 4 suited for independent observation rely on the fact that the set $\left\{d_n(f_{\omega,k}, f_0)^2 + (\sigma - \sigma_0)^2 \geq \epsilon_n^2\right\}$ can be covered with Hellinger balls. Because of the unknown variance, this cannot be done here, we thus use an alternative approach and to construct tests, and then apply Theorem 3 from Ghosal and van der Vaart (2007).

Consider the sets $\mathcal{F}_j^k = \left\{f_{\omega,k}, \sigma; (j\epsilon_n)^2 \leq d_n(f_{\omega,k}, f_0)^2 + (\sigma - \sigma_0)^2 \leq ((j+1)\epsilon_n)^2\right\}$. There exists a constant $C > 0$ such that

$$\mathcal{F}_j^k \subset \left\{||\omega - \omega^0||_n \leq 2j\epsilon_n, |\sigma - \sigma^0| \leq 2j\epsilon_n\right\} \quad (17)$$

To apply Theorem 1 of Ghosal and van der Vaart (2007), we construct test following Choi and Schervish (2007).

For $|\sigma - \sigma_0| \leq \sigma_0/2$. Simple algebra leads to an equivalence between $\left(d_n(f, f')^2 + (\sigma - \sigma')^2\right)^{1/2}$ and the Hellinger metric so that we can apply Lemma 2 of Ghosal and van der Vaart (2007). Equation (17) implies that for all $\xi > 0$ there exist a $\xi\epsilon_n$ net of $\mathcal{F}_j^k$ containing less than $(Dj/\xi)^k$ point with $D > 0$. We then have a test Ψ_1 such that

$$\mathrm{E}_0^n(\Psi_1) \leq e^{-j^2 n\epsilon_n^2}; \sup_{\mathcal{F}_j^k\cap\{|\sigma-\sigma_0|\leq\sigma_0/2\}} \mathrm{E}_{f,\sigma}(1 - \Psi_1) \leq e^{-j^2 n\epsilon_n^2}.$$

For $\sigma > 3\sigma_0/2$ we consider the test Ψ_2 defined as

$$\Psi_2 = \mathbb{I}\left\{\sum_{i=1}^{n}\left(\frac{Y_i - f_0(x_i)}{\sigma_0}\right)^2 > nc_1\right\}$$

for a suitably choosen constant $c_1 > 0$. Chernoff bound gives

$$\mathrm{E}_0^n(\Psi_2) \leq e^{-Cn}$$

for some constant $C > 0$. Note that if $\sigma > 3\sigma_0/2$ and $(f, \sigma) \in \mathcal{F}_j^k$, thus $j > j_0/\epsilon_n$ for some $j_0 > 0$. If $Y_i = f(x_i) + \sigma\varepsilon_i$ where $\varepsilon_i \sim \mathcal{N}(0,1)$ then $\sum_{i=1}^n \left(\frac{Y_i - f_0(x_i)}{\sigma_0}\right)^2$ follow a non central χ_n^2 distribution with non centrality parameter $\sum_{i=1}^n (f(x_i) - f_0(x_i))^2/\sigma^2 > 0$. Thus setting $W \sim \chi_n^2$

$$\mathrm{E}_{f,\sigma}(1 - \Psi_2) = P_{f,\sigma}\left(\frac{\sigma^2}{\sigma_0^2}\sum_{i=1}^n \left(\frac{Y_i - f_0(x_i)}{\sigma}\right)^2 \leq nc_1\right) \leq P_W\left(W \leq cn\frac{\sigma_0^2}{\sigma}\right)$$

Chernoff bound gives

$$E_{f,\sigma}(1 - \Psi_2) \leq e^{-C_2 n}$$

For $\sigma < \sigma_0/2$ we consider the test Ψ_3^t associated to $f^t \in \mathcal{F}_j^k$ a point in the $j\xi\epsilon_n$ net and some suitably choosen $0 < c_2 < 1$ defined as

$$\Psi_3^t = \mathbb{I}\left\{\sum_{i=1}^n \left(\frac{Y_i - f^t(x_i)}{\sigma_0}\right)^2 \leq c_2 n\right\}$$

Similarly to before, given that under P_{f_0,σ_0}, $\sum_{i=1}^n \left(\frac{Y_i - f^t(x_i)}{\sigma_0}\right)^2$ follows a non central χ^2 distribution

$$\mathrm{E}_0^n(\Psi_3^t) = P_0\left(\sum_{i=1}^n \left(\frac{Y_i - f^t(x_i)}{\sigma_0}\right)^2 \leq c_2 n\right) \leq P_W(W \leq c_2 n)$$

Given that the moment generating function of a non central χ_n^2 distribution with non centrality parameter Δ at point s is known to be $(1-2s)^{n/2}\exp\{s\Delta^2/(1-2s)\}$, we have for all $f, \sigma \in \mathcal{F}_j^k \cap \{\sigma < \sigma_0/2\}$ such that $d_n(f^t, f) \leq \xi\epsilon_n$

$$\begin{aligned} &P_{f,\sigma}\left(\frac{\sigma^2}{\sigma_0^2}\sum_{i=1}^n \left(\frac{Y_i - f^t(x_i)}{\sigma}\right)^2 \geq c_2 n\right) \\ &\qquad \leq \exp\left\{\frac{n}{2}\left(-\log(1-2s) + \frac{1}{\sigma^2}\frac{2s}{1-2s}d_n(f, f^t)^2 - 2sc_2\frac{\sigma_0^2}{\sigma^2}\right)\right\} \end{aligned}$$

For s small enough we have

$$\frac{2s}{1-2s}d_n(f, f^t)^2 \leq 4sd_n(f, f^t)^2 \leq 4s\xi^2\epsilon_n^2 \leq 2sc_2\frac{\sigma_0^2}{\sigma^2}.$$

Which in turns gives

$$\mathrm{E}_{f,\sigma}(1 - \Psi_3^t) \leq e^{-nc_2'}$$

Taking $\Psi_3 = \max_t \Psi_3^t$ we get a test such that

$$\mathrm{E}_0^n(\Psi_3) = o(1); \quad \sup_{\mathcal{F}_n^j \cap \{\sigma \leq \sigma_0/2\}} \mathrm{E}_{f,\sigma}(1-\Psi_3) \leq e^{-j^2 n \epsilon_n^2}$$

We conclude the proof by taking $\phi_n = \max\{\Psi_1, \Psi_2, \Psi_3\}$ as an exponentially consistent sequence of tests and applying Theorem 3 of Ghosal and van der Vaart (2007).

B.2 Proof of lemma 2

Denoting $A_n = \left\{d_n^2(f_{\omega,k} - f_0) + |\sigma - \sigma_0|^2 \leq \epsilon_n^2\right\}$ for ϵ_n as in Lemma 1, we have

$$\begin{aligned} &\pi\left(\max_i |\omega_i - \omega_i^0| \geq \xi_n^k |Y^n\right) \leq \\ &\sum_{k=1}^{\infty} \pi(k|Y_n)\pi(\{\max_i |\omega_i - \omega_i^0| \geq \xi_n^k\} \cap A_n | Y^n, k) + o_{P_n^0}(1) \end{aligned} \tag{18}$$

Let

$$\pi(\{\max_i |\omega_i - \omega_i^0| \geq \xi_n^k\} \cap A_n | Y^n, k) = \frac{N_n^k}{D_n^k} = \frac{\int_{\max|\omega_i - \omega_i^0| \geq \xi_n^k \cap A_n} \frac{p(\omega,\sigma,k)}{p(\omega^0,\sigma_0,k)}(Y^n) d\pi(\omega,\sigma)}{\int \frac{p(\omega,\sigma,k)}{p(\omega^0,\sigma_0,k)}(Y^n) d\pi(\omega,\sigma)}$$

For a fixed k, let

$$\begin{aligned} KL^*(f_{\omega,k}, f_{\omega^0,k}) &= \frac{1}{n}\sum_{i=1}^{n} P_{0,i}\left(\log\left(\frac{p_i(\omega,\sigma,k)}{p_i(\omega^0,\sigma^0,k}\right)\right) \\ V^*(f_{\omega,k} f_{\omega^0,k}) &= \frac{1}{n}\sum_{i=1}^{n} P_{0,i}\Bigg(\log\left(\frac{p_i(\omega,\sigma,k)}{p_i(\omega^0,\sigma^0,k}\right) - P_{0,i}\left(\log\left(\frac{p_i(\omega,\sigma,k)}{p_i(\omega^0,\sigma^0,k}\right)\right)\Bigg)^2 \end{aligned}$$

Note that

$$\begin{aligned} KL^*(f_{\omega,k}, f_{\omega^0,k}) &= \frac{1}{2}\log(\sigma_0^2/\sigma^2) - \frac{1}{2}(1 - \frac{\sigma_0^2}{\sigma^2})\left(1 + \frac{1}{\sigma_0^2}(n^{-1}\sum f(x_i)^2 - \omega_0^2\right) - \\ &\quad \frac{1}{2k\sigma^2}\sum_{i=1}^{k}(\omega_i - \omega_i^0)^2 \\ V^*(f_{\omega,k}, f_{\omega^0,k}) &= \frac{1}{2}\left(\frac{\sigma_0^2}{\sigma^2} - 1\right)^2 + \frac{1}{\sigma_0^2}\left(\frac{\sigma_0^2}{\sigma^2} - 1\right)^2\left(n^{-1}\sum f(x_i)^2 - \omega_0^2\right) + \\ &\quad \frac{1}{k\sigma^2}\sum_{i=1}^{k}\left(\frac{\sigma_0^2}{\sigma^2}(\omega_i - \omega_i^0)\right)^2 \end{aligned}$$

Denoting $B_n^*(\xi_n^k) = \left\{KL^*(f_{\omega,k}, f_{\omega^0,k}) \leq (\xi_n^k)^2, V^*(f_{\omega,k}, f_{\omega^0,k}) \leq (\xi_n^k)^4)\right\}$ we have that $B_n^*(\epsilon_n) \supset \{d_n(f_{\omega,k}, f_{\omega_0,k})^2 \leq C(\xi_n^k)^4, |\sigma - \sigma_0|^2 \leq (\xi_n^k)^4\}$, and deduce

$$\pi(B_n^*(\xi_n^k)) \geq e^{-C_d n(\xi_n^k)^2}. \tag{19}$$

Following the proof of Lemma 10 of Ghosal and van der Vaart (2007) we get that

$$P_0^n(D_n^k \geq e^{-(C+1)n(\xi_n^k)^2}) \leq \frac{1}{C^2 n} \tag{20}$$

We now derive an upper bound for N_n^k for a fixed k with high probability. Following Kleijn and van der Vaart (2006) we construct a test ϕ_n^k between the measures P_0^n and $Q_{\omega,\sigma,k}$ defined as $dQ_{\omega,\sigma,k} = p_0/p(\omega^0, \sigma_0, k)dP_{\omega,\sigma,k}$ such that

$$\begin{aligned} P_0\phi_n^k &\leq e^{-C_1 n(\xi_n^k)^2} \\ \sup_{(\omega,\sigma),\max|\omega_j - \omega_j^0| \geq \tau_n^k} Q_{\omega,\sigma,k}(1-\phi_n^k) &\leq e^{-C_2 n(\xi_n^k)^2 + An\epsilon_n^2} \end{aligned} \tag{21}$$

The construction of such test is commonly used to derive convergence rate of the posterior distribution, and general results are known when P_0 and Q are probability measures (see Le Cam, 1986, Chapter 16.4). In the case of Gaussian regression, we compute

$$\begin{aligned} dQ_{\omega,\sigma,k}(Y^n) = & e^{\frac{1}{R}{}_0 2\sigma_0^2 n\left(\frac{\sigma^2}{\sigma_0^2}-1\right)\left(n^{-1}\sum_{i=1}^n f_0(x_i)^2 - \sum_{j=1}^k \omega_j^2\right)} \\ & \prod_{j=1}^k \prod_{x_i \in I_j} \frac{e^{\frac{1}{2\sigma^2}\left(Y_i - \frac{\sigma^2}{\sigma_0^2}(f_0(x_i) - \omega_j^0) - \omega_j\right)}}{\sqrt{2\pi\sigma^2}} \end{aligned} \tag{22}$$

Note that for $(\omega, \sigma, k) \in A_n$ the first term in (22) is bounded by $e^{An\epsilon_n^2}$ where A is a positive constant that may depend on f_0. We thus fall into the usual setting of testing between two probability measures. We thus use the same approach as in Ghosal and van der Vaart (2007) to construct a test that satisfy (21). We first get an upper bound for the metric entropy

$$N(\xi_n^k) = N\left(\xi_n^k/36, \left\{(\omega,\sigma) \in A_n, \max_{i \leq k} |\omega_i - \omega_i^0| \leq \xi_n^k\right\}, h_n\right).$$

Note that

$$\left\{(\omega,\sigma) \in A_n, \max_{i \leq k} |\omega_i - \omega_i^0| \xi_n^k\right\} \subset \left\{d_n(f_\omega, f\omega^0) \lesssim \xi_n^k, |\sigma - \sigma_0| \lesssim \epsilon_n\right\} \tag{23}$$

We can thus derive an upper bound for $N(\xi_n^k) \lesssim (\xi_n^k)^{-k}$. We thus build a test ϕ_n^k satisfying (21) in the same way as in Ghosal and van der Vaart (2007). We thus compute

$$
\begin{aligned}
P_0^n\left(\frac{N_n^k}{D_n^k}\mathbb{I}_{A_n} \geq 1/2\right) &\leq 2\mathrm{E}_0^n\left(\frac{N_n^k}{D_n^k}\mathbb{I}_{A_n}\right)\\
&\leq 2\mathrm{E}_0^n\left(\frac{N_n^k}{D_n^k}\mathbb{I}_{A_n}(1-\phi_n^k)\mathbb{I}_{D_n^k > e^{-Ck\log(n)}}\right) + \frac{2C}{n} + 2e^{-C_2 n(\xi_n^k)^2}\\
&\lesssim e^{-Cn(\xi_n^k)^2}\mathrm{E}_0^n\left(N_n^k(1-\phi_n)\right) + \frac{2C}{n} + 2e^{-C_2 n(\xi_n^k)^2}\\
&\lesssim e^{-Cn(\xi_n^k)^2}\mathrm{E}_0^n\left(\mathbb{I}_{A_n}\int_{\max|\omega_i-\omega_i^0|>c\tau_n^k} Q_{\omega\sigma,k}(1-\phi_n)d\pi(\omega,\sigma)\right) +\\
&\frac{2C}{n} + 2e^{-C_2 n(\xi_n^k)^2}\\
&\lesssim e^{-Cn(\xi_n^k)^2 + An\epsilon_n^2} + \frac{C}{n} \qquad (24)
\end{aligned}
$$

If $f_0 \in \mathcal{F}$, we have $n\epsilon_n^2 \approx \log(n)$ and thus the first term of (24) is less than $e^{-Ck\log(n)}$. We thus have

$$
P_0^n\left(\frac{N_n^k}{D_n^k}\mathbb{I}_{A_n} \geq 1/2\right) \lesssim \frac{1}{n}
$$

Given (18) we deduce (13) for $f_0 \in \mathcal{F}$. If $f_0 \notin \mathcal{F}$ and $f_0 \in H_\alpha(L)$ we have that for M large enough $An\epsilon_n^2 \leq C\rho_n(\alpha)/2$, then the first term in (24) is less that $e^{-C\rho_n(\alpha)/2}$ and thus

$$
P_0^n\left(\frac{N_n^k}{D_n^k}\mathbb{I}_{A_n} \geq 1/2\right) \lesssim \frac{1}{n}
$$

which gives (13).

B.3 Proof of Lemma 3

Let f_0 be in $H_\alpha(L)$ and $k_n = (n/\log(n))^{1/(2\alpha+1)}$. Similarly to before, we have $\pi\left(B_n(\epsilon_n)\right) \geq e^{-n\epsilon_n^2}$. We define N_n and D_n such that

$$
\pi(\mathcal{K}_n^c|Y_n) = \frac{\sum_{k\in\mathcal{K}_n^c}\pi(k)\int\frac{p(\omega,\sigma,k)}{p_0}(Y^n)d\pi(\omega,\sigma)}{\sum_k\pi(k)\int\frac{p(\omega,\sigma,k)}{p_0}(Y^n)d\pi(\omega,\sigma)} = \frac{N_n}{D_n}
$$

Given Lemma 10 of Ghosal and van der Vaart (2007), we have

$$
P_0^n\left(D_n \leq e^{-Cn\epsilon_n^2}\right) = o(1)
$$

Note also that

$$
\mathrm{E}_0^n(N_n) = \sum_{k\in\mathcal{K}_n^c}\pi(k)\int\int_{\mathbb{R}^n}\frac{p(\omega,\sigma,k)}{p_0}(Y^n)p_0(Y_n)d\pi(\omega,\sigma)dY^n = \pi(k\leq k_n) \leq ce^{-C_u k_n L(k_n)}
$$

Thus for C small enough we have

$$\begin{aligned}\mathrm{E}_0^n\left[\pi\left(k \in \mathcal{K}_n^c | Y^n\right)\right] &= \mathrm{E}_0^n\left[\frac{N_n}{D_n}\mathbb{I}_{D_n > e^{-Cn\epsilon_n^2}}\right] + o(1)\\ &\leq e^{Cn\epsilon_n^2} c e^{-C_u k_n L(k_n)} + o(1)\\ &\leq o(1)\end{aligned}$$